The industrial melanism mutation in British peppered moths is a transposable element

Arjen E van't Hof[1]*, Pascal Campagne[1]*, Daniel J Rigden[1], Carl J Yung[1], Jessica Lingley[1], Michael A Quail[2], Neil Hall[1], Alistair C Darby[1], Ilik J Saccheri[1]

[1] Institute of Integrative Biology, University of Liverpool, Biosciences Building, Crown Street, Liverpool L69 7ZB, UK.
[2] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK.
* These authors contributed equally to this work.

Discovering the mutational events that fuel adaptation to environmental change remains an important challenge for evolutionary biology. The classroom example of a visible evolutionary response is industrial melanism in the peppered moth (*Biston betularia*): the replacement, during the Industrial Revolution, of the common pale *typica* form by a previously unknown black (*carbonaria*) form, driven by the interaction between bird predation and coal pollution[1]. The *carbonaria* locus has been coarsely localised to a 200 kilobase region but the specific identity and nature of the sequence difference controlling the *carbonaria-typica* polymorphism, and the gene it influences, are unknown[2]. Here we show that the mutation event giving rise to industrial melanism in Britain was the insertion of a large, tandemly repeated, transposable element (TE) into the first intron of the gene *cortex*. Statistical inference based on the distribution of recombined *carbonaria* haplotypes indicates that this transposition event occurred around 1819, consistent with the historical record. We have begun to dissect the mode of action of the *carbonaria*-TE by showing that it increases the abundance of a *cortex* transcript, whose protein product plays an important role in cell-cycle regulation, during early wing disc development. Our findings fill a significant knowledge gap in the iconic example of microevolutionary change, adding a further layer of insight into the mechanism of adaptation in response to natural selection. The discovery that the mutation itself is a TE will stimulate further debate about the importance of 'jumping genes' as a source of major phenotypic novelty[3].

Ecological genetics, the study of polymorphism and fitness in natural populations, has been revitalised through the application of next-generation sequencing technology to open up what were previously treated as genetic black boxes[4,5]. Growing appreciation of the loci and developmental networks that generate adaptive phenotypic variation[6] promises to answer fundamental questions about the genetic architecture of adaptation, such as the prevalence of genomic hotspots for adaptation[7], the relative contribution of major vs minor effect mutations[8], and the structural nature and mode of action of beneficial mutations[9]. The significance of characterising the identity and origin of functional sequence polymorphisms is in providing an explicit link between the mutation process and natural selection. In this context, whilst industrial melanism in the peppered moth has retained its appeal as a graphic example of the spread of a novel mutant rendered favourable by a major change in the environment, the crucial piece of the puzzle that has been missing is the molecular identity of the causal mutation(s)[10].

A combined linkage and association mapping approach previously localised the *carbonaria* locus to a < 400 kb region orthologous to *Bombyx mori* chromosome 17 (loci *b-d*)[2]. Thirteen genes and two miRNAs occur within this interval, none of which were known to be involved in wing pattern development or melanisation. By extending the association mapping approach to a larger population sample and more closely spaced genetic markers (Methods), the *carbonaria* candidate region was narrowed to ~100 kb (Fig. 1a). The candidate region resides entirely within the span of one gene – the ortholog of *Drosophila* '*cortex'* (*cort*), whose only known function is as a cell-cycle regulator during meiosis[11]. In *B. betularia*, *cortex* consists of 8 non-first exons, multiple alternative first exons (of which only two, 1A and 1B, are strongly expressed in developing wing discs), and a very large first intron (Fig. 1b).

The rapid spread of *carbonaria* gave rise to strong linkage disequilibrium (LD)[2], such that many sequence variants are associated with the *carbonaria* phenotype. This poses a challenge for isolating the specific causal variant(s). We reasoned that if the *carbonaria* mutation arose on an ancestral *typica* haplotype[2], the hitchhiking variants should in principle also be present at some frequency within the *typica* population, leaving the causal variants as the only ones unique to *carbonaria*. High quality contiguous reference sequences were assembled from tiled BAC and fosmid clones, resulting in one *carbonaria* and three different *typica* core haplotypes (Methods; Extended Data Fig. 1). Alignment of these sequences (Supplementary Text 1) revealed 87 melanisation candidate polymorphisms (Fig. 1b; Supplementary Table 1), concentrated within the large first intron of *cortex* (69-91 kb, depending on haplotype). Eighty-five candidates were eliminated using an increasing number of *typica* individuals to exclude rare variants. A single nucleotide polymorphism (carbonaria_candidate_25) was eventually excluded on the basis of one individual out of 283 *typica*, leaving a very large insert (carbonaria_candidate_45) as the only remaining candidate.

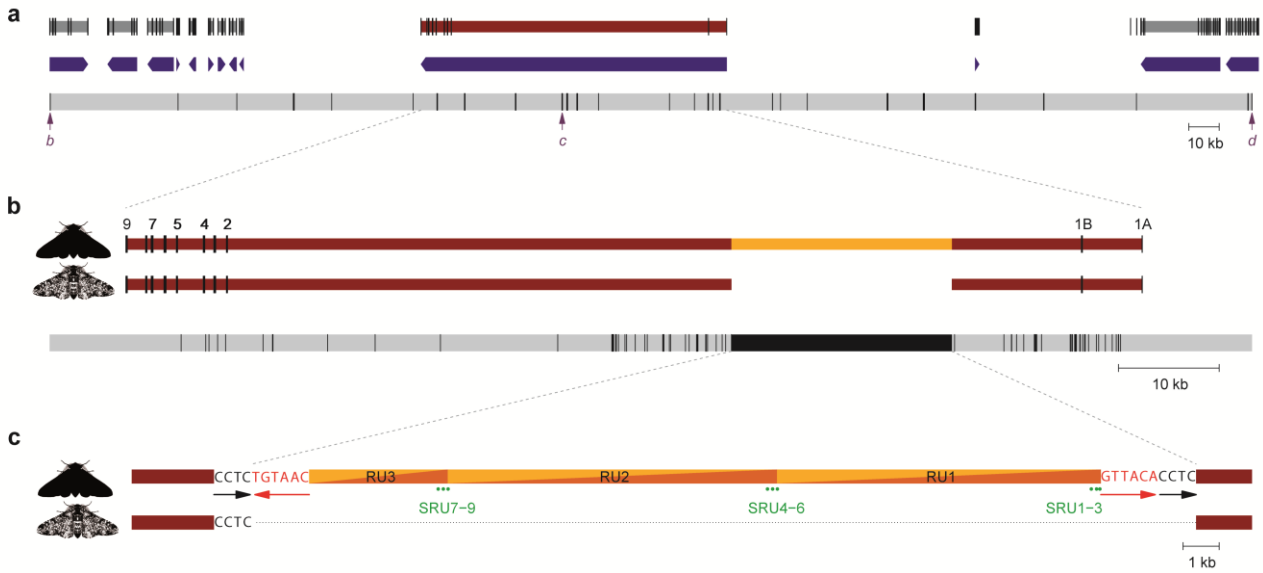


Figure 1: The *carbonaria* candidate region, and the position and structure of the *carbonaria* mutation. a,~400 kb candidate region (bounded by marker loci *b* and *d* [2]) indicating gene content and genotyping positions (vertical lines in the continuous grey bar). Intron-exon structure and orientation are illustrated separately for each gene (annotated in GenBank KT182637). **b**, Refined candidate region including candidate polymorphisms (lines on the grey bar). The intron-exon structure of *cortex* is shown for *carbonaria* (black moth) and *typica* (speckled moth), highlighting the presence of a large (22 kb) indel (orange) within the first intron. Exons 1A and 1B are alternative transcription starts followed by the shared exons 2-9. **c**, The only exclusive *carbonaria-typica* polymorphism within the candidate region. The structure of the insert, shown in the *carbonaria* sequence, corresponds to a class II DNA transposon, with direct repeats resulting from target site duplication (black nucleotides) next to inverted repeats (red nucleotides). *Typica* haplotypes (lower sequence) lack the 4-base target site duplication, the inverted repeats and the core insert sequence. The transposon consists of ~9 kb tandemly repeated 2⅓ times (RU1-RU3), with three short tandem subrepeat units (green dots, SRU1-SRU9) within each RU.

The insert was found to be present in 105 out of 110 fully black moths (wild caught in the UK since 2002) and absent in all (283) *typica* tested (Methods; Extended Data Fig. 2). Consistent with local *carbonaria* morph frequencies of 10-30%[12], 2/105 individuals were homozygous for the *carbonaria* insert. Five individuals morphologically indistinguishable from *carbonaria* did not possess the *carbonaria* insert; they do not present any strong haplotype association based on this set of candidate loci but do all differ from the core *carbonaria* haplotype at many positions. Our interpretation is that these individuals are hetero- or homo-zygous for the most extreme of the *insularia* alleles (intermediate phenotypes), which are known to occasionally produce *carbonaria*-like phenotypes[13,14] and segregate as alleles of the *carbonaria* locus in classical genetics crosses[14]. Conversely, none of the genotyped *insularia* morphs (31 individuals, covering the full spectrum of variation from i_1 to i_3[14]) contain the *carbonaria* insert (Extended Data Fig. 2). We conclude that the large insert is the *carbonaria* mutation.

The *carbonaria* insert is 21,925 nt in length, composed of a ~9 kb essentially non-repetitive sequence (except for ~370 nt at the repeat unit junctions) tandemly repeated approximately two and one-third times, with only minor differences among the repeats (Fig. 1c). The insert bears the hallmark of a class II (DNA cut-and-paste) transposable element: short inverted repeats (6 bp) and duplication of the (4 bp) target site present in *typica* haplotypes (Extended Data Fig. 3). We estimate ~255 and ~60 genomic copies, respectively, of the 9 kb *carbonaria*-TE repeat unit (RU) and RU junctions, implying relatively few genomic copies of the complete *carb*-TE. No nucleotide or translated BLAST hits were found in any relevant database, with the exception of *B. betularia* RNAseq reads (NCBI: SRX371328), indicating that the *carb*-TE RU is *Biston*-specific.

In order to examine patterns of recombination, which provide insight into the evolutionary dynamics of a chromosomal region, we genotyped the same 105 *carbonaria* and a sub-set of 37 *typica*, plus 35 *insularia*, at 119 polymorphic loci within 28 PCR fragments distributed along ~200 kb either side of the *carb*-TE (Fig. 1a). Diploid genotypes were phased, and the resulting haplotypes divided into those with and without the *carb*-TE. The sequence identity of the ancestral *carbonaria* haplotype, whose core was known from the BAC/fosmid work, was extended by assigning allelic state at each marker locus to ancestral *carbonaria* or *typica/insularia*. Fifty percent of *carb*-TE haplotypes have retained the ancestral *carbonaria* haplotype across the full 400 kb window, the remainder showing

varying degrees of recombination with *typica* haplotypes on one or both sides of the causal mutation (Fig. 2a). The recent selective sweep[15] is reflected by declining LD between the *carbonaria* locus and marker loci with increasing genetic distance (Fig. 2b). Tenure of the *carb*-TE has been transient, having declined from ~99% to less than 5% in its industrial heartland since 1970[16]. It has nevertheless left a substantial trace of its former abundance in the form of ancestral *carbonaria* haplotype blocks introgressed into *typica* and *insularia* haplotypes (Fig. 2c), consistent with the simulation-based expectation.

1848 Manchester is generally regarded as the first reported sighting of the *carbonaria* form[1], although the wording of the record implies that it was rare but not completely unknown at this time. Establishing how long before this date the *carbonaria* mutation occurred is complicated because it could have existed undetected at low frequency for hundreds of years (Supplementary Methods). Our approach to this problem has been to independently infer the age of the mutation event by considering the erosion of the ancestral *carbonaria* haplotype due to genetic recombination and mutation. One million simulated time trajectories of the *carbonaria* phenotype were randomly drawn according to their fit to historical frequency data (Extended Data Fig. 4). Based on these trajectories, recombination patterns were simulated using an empirical estimate of recombination rate and compared to the observed recombination pattern of the *carbonaria* haplotypes. The probability density for the date of the *carb*-TE mutation event (Fig. 2d) is highly skewed (median = 1763, interquartile range = 1681-1806) with a maximum likelihood at 1819, a date highly consistent with a detectable frequency being achieved in the mid-1840s.

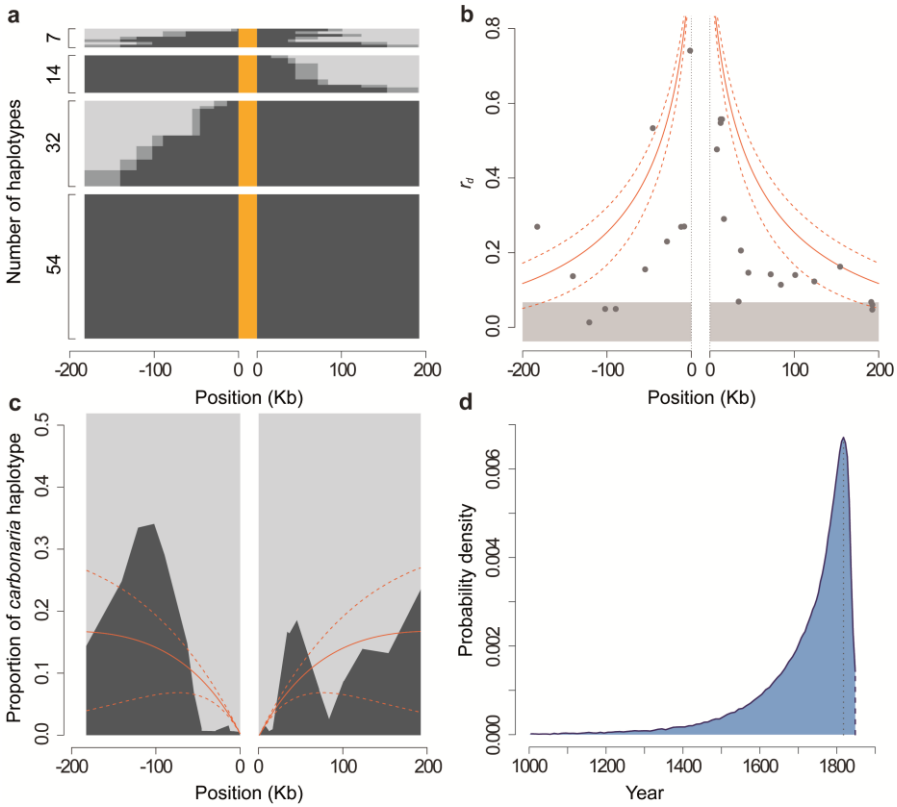


Figure 2: Recombination pattern and ageing of the *carb-TE* mutation. a, Nearest recombination sites ~200 kb either side of the *carb*-TE (at position 0) in a sample of 107 *carbonaria* haplotypes, i.e. *carb*-TE present (orange), with non-*carbonaria* (*typica* and *insularia*) haplotypes (light grey). Dark grey areas indicate boundaries within which recombination occurred. **b**, Multilocus linkage disequilibrium (r_d) across the same sequence window among *carbonaria* and non-*carbonaria* haplotypes. Grey area indicates the widest 99% confidence region, across loci, for the null hypothesis ($r_d \approx 0$). Red lines represent the simulation-based upper bound under the extreme assumption that all alleles defining the *carbonaria* haplotype were initially exclusive to it (mean and 90% interval). **c**, Introgression of the ancestral *carbonaria* haplotype (black) into non-*carbonaria* haplotypes (grey), i.e. *carb*-TE absent (n = 144). Red lines represent the simulation-based expectations (mean and 90% interval). **d**, Probability density for the age of the *carb*-TE mutation inferred from the recombination pattern in the *carbonaria* haplotypes (maximum density at 1819 shown by dotted line; first record of *carbonaria* in 1848 shown by dashed line).

The position of the *carb*-TE suggests that its effect on melanisation is achieved through altering the expression of *cortex*, through one of several potential mechanisms[17] (incorporation of any part of *carb*-TE into *cortex* transcripts has been excluded). *Biston cortex* is characterised by numerous splice isoforms and alternative first exons; we focus on the population of transcripts initiated by exons 1A and 1B, as the other first exons are absent or only weakly expressed in *Biston* wing discs, and did not exhibit morph-specific differences (Extended Data Fig. 5). The global pattern of splice isoforms showed neither consistent presence/absence or crude relative abundance differences among morphs for any developmental stage (Extended Data Fig. 6 and 7). Cumulative expression across all isoforms (Fig. 3a) increases by an order of magnitude between the 6th larval instar (La6) and day 4 prepupa (Cr4), coinciding with a phase of rapid wing disc morphogenesis (Fig. 3b), falling back to a low level by day 6 prepupa (Cr6) – with no clear difference among morphs (*t/t* vs *c/t*, $P > 0.5$). To exclude interference by potentially non-functional isoforms, we targeted full transcripts only, either starting with 1A or 1B. The trend for the abundance of 1B full transcript, consistent across several families with different genetic backgrounds, is *c/c* > *c/t* > *t/t*, most pronounced at Cr4 (Fig. 3c and

Extended Data Fig. 8a). 1A-initiated full transcript, which is in general an order of magnitude less abundant than 1B, does not show a significant difference between genotypes (Fig. 3d and Extended Data Fig. 8b).

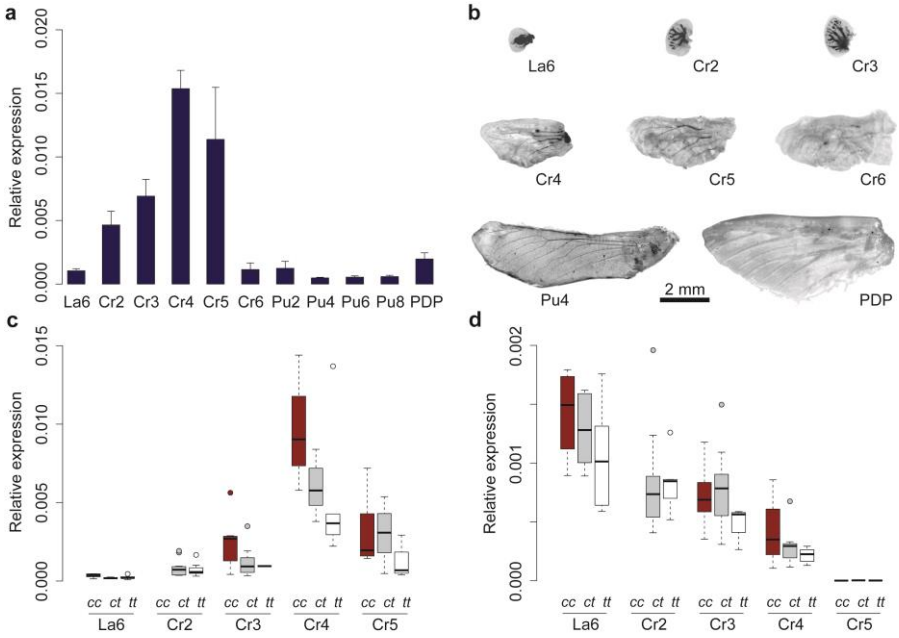


Figure 3: Relative expression of *cortex* in developing wings of *B. betularia*. **a**, Average expression (across *typica* and *carbonaria* morphs) of all *cortex* splice variants (exons 7-9) relative to the control gene *Spectrin alpha chain* in wing discs at different developmental stages (La6: 6th instar larvae, Cr2: day 2 crawler, Pu2: day 2 pupae, PDP: post diapause pupae). **b**, Scaled images of *B. betularia* forewings at different stages. **c** and **d**, Relative expression of *cortex* 1B (**c**) and 1A (**d**) full transcript in developing wings of the three *carbonaria*-locus genotypes (*cc*, *ct* and *tt*) produced within the progeny of a *ct* x *ct* cross (no data for *cc* at Cr2). Genotypes differ significantly for 1B full transcript ($P < 0.001$), whereas genotypes do not differ for 1A full transcript ($P > 0.2$). (Note the differing y-axis scales). Equivalent graphs for the progeny of *ct* x *tt* crosses (which lack the *cc* genotype) are presented in Extended Data Fig. 8.

The role of *cortex* in wing pattern melanisation is not obvious. In *Drosophila*, *cortex* has been primarily associated with meiosis in ovaries[11] (several *cortex* transcripts are expressed in *Biston* ovaries and testes, Extended Data Fig. 5). Molecular function is suggested by phylogenetic analysis which indicates that *Biston cortex* occurs in a lepidopteran sub-group within an insect-specific clade of a protein family containing cell-cycle regulators Cdc20 and Cdh1, encoded by *Fzy* and *Fizzy-related* in *Drosophila* (Extended Data Fig. 9b). These proteins help regulate fundamental cell division processes such as cytokinesis by presenting substrates to, and activating, the anaphase-promoting complex or cyclosome (APC/C), which ubiquitinates cell-cycle proteins, thereby earmarking them for degradation. Substrate recognition is by binding to degrons, short linear motifs such as the D box and KEN box. Sequence conservation across lepidopterans and non-lepidopterans reveals a single binding site in cortex (Extended Data Fig. 9c) which likely binds the D box-like[18] degron LxExxxN[19]. This degron binding capability is predicted for both of the full isoforms (1A [441aa] and 1B [407aa], although 1B apparently lacks the N-terminal C box usually required for APC/C binding) but not for the alternative isoforms (Extended Data Table 1). These data demonstrate orthology and support shared function between cortex in *D. melanogaster* and *B. betularia*, although the molecular connection between cell-cycle protein degradation at the APC/C and melanisation remains to be determined.

Our results suggest that *carb*-TE impacts adult melanisation pattern through increasing the abundance of cortex, perhaps altering the course of scale cell heterochrony, with dominance arising through a threshold effect (1B full transcript is more abundant in *c/c* than *c/t*). How the *carb*-TE promotes *cortex* expression is unknown but the general mechanism is predicted to allow for the production of *insularia* morphs putatively controlled by different mutations within *cortex*. In combination with parallel findings in *Heliconius* butterflies[20], our results support the idea that *cortex* is a conserved developmental node for generating colour pattern variation in evolutionarily diverse Lepidoptera. It may, however, not be the only gene in this region involved in patterning, as suggested by recent work on *Bombyx mori* mutant, *Black moth*, which has a similar phenotype to *B. betularia carbonaria*[21], although none of the genes implicated are differentially expressed among *carbonaria* and *typica* wing discs.

The *carb*-TE is a spectacular example of an adaptively advantageous transposon[22-24], its discovery filling a fundamental gap in the peppered moth story and furthering our appreciation of the mechanism underpinning rapid adaptation. A consensus on the general importance of TEs for adaptive evolution has yet to emerge[3,25]. Over longer time frames, phenotypic effects of TEs may be obscured by imprecise excision that leaves a minimal trace of the TE whilst retaining the mutant (adaptive) phenotype[26]. By contrast, we have shown that the *carb*-TE is young, approximately 200 years (generations) old, during which time it has gone from a single mutation to near fixation (regionally) to near extinction – driven by a pulse of environmental change.

Methods

Wild samples

Moths used for fine mapping and ageing analysis came from a northwest England – north Wales transect sampled in 2002 [12], with 12 *carbonaria* and 6 *insularia* specimens additionally collected in 2005-2009.

Reference sequences

An extended BAC tiling path was constructed using mapped *B. betularia* genes [27], *B. mori* nscaf2829 (SilkDB) orthologs, and BAC-end sequences as probes. Combinatorial PCR using BAC-end sequences and internal gene anchors were used to determine the relative positions of the BACs. Fosmids were used to bridge a gap. A minimal tiling path was sequenced as a 3 kb mate-pair library with Roche 454 GS FLX Titanium. Reads were assembled into contigs using Newbler and manually scaffolded using tiled BAC-end sequences and exon order of genes spanning multiple contigs as anchors. The scaffold covers a 3.6 Mb region spanning from the mapped genes *myosin heavy chain* (*myosin HC*) to *leucine-rich transmembrane protein* (*LRTP*) with the *carbonaria* polymorphism located towards the centre. A recombination rate estimate within this region of 2.9 cM/Mb was obtained from a total of 350 offspring in 8 crosses screened for recombination between the ends of the 3.6 Mb interval. Three *typica* and one *carbonaria* haplotype sequences were reconstructed using BACs and fosmids for the region spanning locus *b* to *d* (Fig. 1a). Clones were assigned to haplotypes based on co-segregation of genotypes and phenotypes between parents and sibs of the heterozygous (*carbonaria*-*typica*) individuals used to generate the BAC (family 67) and fosmid (family 11) libraries. Small assembly gaps caused by repetitive sections were bridged using long capillary Sanger sequences; fosmid clone 25H14, containing the large repetitive TE, was sequenced using Pacific Biosystems RS II to 300x coverage using P4-C2 chemistry and assembled using HGAP v2 (Pacific Biosystems). Homopolymer length variation often caused by 454 errors rather than true polymorphisms was verified by Sanger sequencing.

A draft genome assembly was generated from an individual homozygous for the *carbonaria* region. Full-sib *carbonaria*-*typica* heterozygotes were crossed (family 135) to produce homozygous *carbonaria* offspring, as well as heterozygotes and homozygous *typica*. The *carbonaria* homozygotes were identified using alleles closely linked to the *carbonaria* locus, with more distant loci on either side used to ensure that the haplotype had not been disrupted by recombination. DNA was prepared by phenol-chloroform extraction from a final instar male larva with gut removed. The genome was sequenced at ~3.5X coverage on a 454 FLX+ platform and a draft assembly constructed using Newbler. The genome assembly was used for polymorphism discovery, and in tiling path construction using homology to *B. mori*. Single read coverage was used to detect repetitive regions, aiding in single-target primer design, and to confirm the repetitive nature of the *carb*-TE.

The gene content of the *b*-*d* interval was examined by comparing its sequence against GenBank proteins, ESTs, transcriptomes, and annotated genes in the orthologous region in other Lepidoptera. Tblastx against these orthologous regions and Augustus [28] gene prediction were used to detect potentially overlooked genes. All genes were manually annotated and (except for *vcpl*) confirmed using cDNA. The annotation of 11 genes (not including *cortex*) was also subsequently confirmed against a *B. betularia* transcriptome (GenBank # SRX371328) assembled with Trinity [29]. MicroRNAs were found using miRBase with blastn including hairpin precursors. BLAST (blastn, blastx) searches for *carb*-TE-like sequences were performed on NCBI databases (GenBank nt, protein, EST, transcriptome), independently curated lepidopteran genome assemblies (e.g. SilkDB), and RepBase (19.09).

Fine mapping

The interval containing the *carbonaria* polymorphism was narrowed down to a section bordered on both sides by evidence of *carbonaria* haplotype breakdown caused by recombination. Polymorphisms at regular intervals in the *b*-*d* region (Fig. 1a; Supplementary Table 2) were genotyped in wild-caught *carbonaria*, *typica* and *insularia* (105, 33 and 30 individuals, respectively). We conservatively used only homozygous genotypes to set these boundaries since the dominance of *carbonaria* obscures the assignment of alleles in heterozygous genotypes to a certain morph haplotype. The four contiguous haplotype sequences (one *carbonaria*, three *typica*) constructed from BACs and fosmids were aligned between these narrowed down boundaries and examined for polymorphisms that were distinct in the *carbonaria* haplotype relative to all three *typica* haplotypes, resulting in 87 *carbonaria* candidate polymorphisms (Extended Data Fig. 1, Supplementary Table 1). With the exception of carbonaria_candidate_45, wild-caught *typica* were genotyped at all loci by means of PCR-RFLP, PCR-indel or sequencing. Dependent on the frequency of the candidate alleles in the *typica* sample, 16 to 283 *typica* (32 to 566 *typica* haplotypes) were used for exclusion. Carbonaria_candidate_25 was present in only one out of 566 *typica* haplotypes. The *typica* phenotype of this

individual (12-2002-01) was confirmed, as was the presence of carbonaria_candidate_25 allele from independently extracted DNA. A very large indel, later identified as the true *carbonaria* polymorphism (carbonaria_candidate_45), that could not be bridged by PCR required an alternative present/absent screening approach which also provides a positive control for absence haplotypes (to distinguish insert absence from PCR failure). A three-primer PCR was designed with two primers flanking the indel and a third within the insert, relatively close to the indel boundary (Extended Data Fig. 2). The assay was validated using a family known to include all three genotypes (Family 135, Extended Data Fig. 2).

Inferring haplotypes and the age of the *carbonaria* mutation

A set of 177 individuals, including 105 *carbonaria* individuals was genotyped at 119 polymorphic loci within 28 PCR products, stretching across ~400 kb (Supplementary Table 2). *Carbonaria* haplotypes were inferred using SHAPEIT [30] and the position (interval) of recombination breakpoints inferred based on two or more consecutive phase-switched polymorphisms. High repeatability of the phasing outcomes was verified by resampling, and switch errors minimised by including known haplotypes and classifying only two types (melanic and non-melanic). Indices of multilocus linkage disequilibrium (r_d) were calculated on polymorphisms within each PCR fragment and the *carbonaria* locus across the 400 kb interval [31]. Their significance was assessed using 999 Monte-Carlo permutations. The pattern of introgression of the *carbonaria* haplotype into background haplotypes (i.e., *typica* and *insularia* morph alleles) was assessed using ChromoPainter v2 [32] to search for contiguous blocks that match the *carbonaria* haplotype, thus generating the 'expectation painting' of background haplotypes.

The age of the *carbonaria* mutation was inferred with a simulation-based approach. The analysis was performed in three steps. Firstly, 1,000,000 time-forward trajectories of the *carbonaria* phenotype were sampled, using a Metropolis-Hastings algorithm, depending on their likelihood given historical phenotypic frequencies (Supplementary Table 3), and conditional to their starting date (x_0) and population size (N). Secondly, recombination patterns were simulated using the sampled trajectories, in populations of size N, and a fixed recombination rate of 2.9 cM/Mb (males only). This process yielded sample distributions of the closest recombination breakpoint relative to the *carbonaria* locus. Finally, the likelihood of the simulated distributions given the empirical recombination pattern was computed and averaged out across simulations in order to estimate the probability density of the mutation age (x_0). Full details in Supplementary Methods.

Expression and alternative transcripts of *cortex*

Offspring from either heterozygous *carbonaria*/*typica* (*c*/*t*) x homozygous *t*/*t* crosses segregating 1:1, or *c*/*t* x *c*/*t* crosses segregating 1 *c*/*c*: 2 *c*/*t*: 1 *t*/*t*, were used for end-point reverse transcription PCR (RT-PCR) and real time qPCR (RT-qPCR) experiments. Caterpillars were reared on grey willow (*Salix cinerea*). Wing discs (forewings and hindwings) were dissected from final (6^{th}) instar larvae, crawlers or prepupae (day 2-6 from the start of crawling stage), pre-diapause pupae (day 2-8 from pupation, at which point they have entered diapause) and post-diapause pupae (wing discs staged into six categories), and stored in RNAlater (Ambion). RNA was extracted with TRIzol and cDNA synthesized with SuperScript III (Invitrogen) – oligo(dT). The genotype-phenotype (adult morph) of each wing disc specimen was determined with the *carb*-TE three primer PCR (and verified by sequencing a linked SNP, carbonaria_candidate_25). Relative abundance and RT-qPCR data were analysed using Generalized Linear (mixed) Models. See Extended Data Fig. 8c for sample sizes.

RT-qPCR experiments were designed to measure the relative abundance of *cortex* transcripts, either of all transcripts combined (using primers in exons 7 and 9) or full transcripts only (primers in exons 1A-3 and 1B-3, as exon 3 is effectively exclusive to the full transcripts [Extended Data Fig. 7]). DNAse treatment was not performed, but for exons 7-9 RT-qPCR co-amplification of gDNA was prevented by positioning the reverse primer on the exon 8-9 boundary (this was not a concern for exons 1-3 RT-qPCRs because the large first intron precluded gDNA amplification). We chose *40S ribosomal protein S3a* [33] and *Spectrin alpha chain* [34] as two single-copy autosomal housekeeping genes. Primer sequences are listed in Supplementary Table 4. Annealing temperatures were optimised to 66°C and amplicons were confirmed to produce single bands on agarose gels. cDNA was diluted 1:1 with water to allow template volumes within the accuracy range of the pipette used. RT-qPCRs for target and control were run in three replicates using Kapa SYBR Fast qPCR Universal under recommended conditions on a Roche LightCycler 480 with 45 cycles and a melting curve. Since both control genes gave similar results, only *Spectrin alpha chain* was used for the entire sample.

Alternative transcription starts of *cortex* were searched for using 5' RACE on RNA extracted from 15 wing disc samples covering a wide range of stages and *c/c, c/t, t/t* genotypes, and additionally from whole pupa and testes. *Cortex*-specific cDNA was synthesised with SuperScript III and a gene-specific negative strand primer; 5' cytosine extension was added using terminal transferase (NEB) and dCTPs. The single-stranded cDNA was made double-stranded and a target sequence for amplification incorporated in a single extension cycle (LongAmp Hot Start, NEB) with an oligo containing a 5' primer recognition site and a 3' poly-G tail. PCR was performed using a forward primer matching the synthetic 5' end and a nested *cortex*-specific reverse. The amplicons were sequenced using a second nested primer. The alternative first exons were confirmed by Sanger sequencing with forward primers inside the new-found exons to generate clean sequence without the background noise commonly observed with 5' RACE.

The complete pattern of *cortex* splice variation was examined with end-point RT-PCR using primers Bb_cort_exon1A_F, or Bb_cort_exon1B_F, and Bb_cort_exon9_R (primer sequences in Supplementary Table 4). PCR conditions were 60°C annealing, 40 cycles, 75 s extension, 25 µl total volume, 3 µl wing disc cDNA, LongAmp Taq DNA polymerase (NEB). A Fragment Analyzer (Advanced Analytical) was used to estimate the size and relative abundance of amplicons within each individual, after normalising samples to a concentration range of ~1-10 ng/µl. The concentration of each fragment peak was calculated using PROSize (Advanced Analytical), and the relative abundance was computed as the concentration of a splice variant divided by the sum of all fragment concentrations within that individual profile. The *cortex* splice variant amplicons were sequenced as two pools (*t/t* and *c/t*) using Pacific Biosystems RS II with P6-C4 chemistry and the insert reads extracted using smrtportal (Pacific Biosystems). Reads that contained exon 1A or 1B and exon 9 were used to validate the sequence composition and relative abundance of spliced gene isoforms.

No part of *carb*-TE was detected in *cortex* transcripts, either with PacBio sequencing or with PCR using various primer combinations where one primer lies within the transposon and the other matches a *cortex* exon. However, *carb*-TE-like partial sequence was amplified (with primers within repeat units) from both *typica* and *carbonaria* morph cDNA synthesised using *carb*-TE primers, implying that these RNA sequences are transcribed from non-allelic homologs of the *carb*-TE.

Expression of alternative candidate genes

Two *Bombyx mori* adult melanism/patterning mutants, *Black moth* (*Bm*) and *Wild wing spot* (*Ws*), were recently mapped to a region partially orthologous to the *carbonaria* interval [21]. In this study, end-point PCR showed complete absence of *cortex* expression in pupal stages and adults while potentially important prepupal stages were not examined. Three neighbouring genes (BGIBMGA005658, 5657, 5655) did show convincing differences between wild type and both mutants even though these genes lie outside the *Ws* mapping interval. We performed equivalent end-point reverse transcription PCR in *B. betularia* for the three orthologs *gloverin-2, menm* and *lrrp* to determine whether morph – gene expression associations existed between *carbonaria* and *typica* (comparing *c/t* and *t/t* genotypes for wing disc stages: Cr4, Cr6, Pu2, Pu4 and PDP). PCR conditions as for *cortex* 1A/1B-9 end-point PCRs, except 45 seconds extension (primer sequences in Supplementary Table 4).

Cortex phylogeny and protein modelling

Cortex sequences derived from database searches (Supplementary Table 5) were supplemented with a selection of Chd1/Cdc20/Fizzy sequences from model organisms and the set aligned with MAFFT [35] (Supplementary Text 2). The central propeller domain was isolated and used for bootstrapped phylogenetic analysis with MEGA 6 [36] employing its Maximum Likelihood algorithm and the JTT matrix-based model. Any gapped positions were ignored. Homology models of *B. betularia* and *D. melanogaster* cortex proteins were made with MODELLER [37] and Consurf [38] used to map protein sequence conservation to their respective surfaces, among lepidopteran or non-lepidopteran cortex proteins.

Accession codes

Typica 1 haplotype (*b-d* interval) reference sequence (KT182637); *Biston betularia* whole genome sequence NCBI SRA database (SRX1060177-SRX1060182); *cortex* splice variants (KT235895-KT235906); *Rps3A* (JF811439); *α-Spec* (KT182638).

Acknowledgements

University of Liverpool – Centre for Genomic Research (Margaret Hughes, Christian Bourne, Richie Eccles, Christiane Hertz-Fowler, John Kenny) performed next-gen sequencing and Fragment Analyzer measurements.

Laurence Cook directed us to historical data sources, Casey Bergman advised on transposon detection, four anonymous reviewers provided valuable comments on an earlier version of this ms. Population genetics simulations were performed on the University of Liverpool Advanced Research Computing Condor service. This work was supported by Natural Environment Research Council grants NE/H024352/1 and NE/J022993/1.

Author contributions

I.J.S., A.E.H. and P.C. designed the study and wrote the paper (P.C., A.E.H. and D.J.R. produced the figures); A.E.H. directed molecular biology experiments; A.E.H., C.J.Y. and J.L. conducted molecular biology experiments; A.E.H. constructed the BAC and fosmid tilepaths; A.E.H. and A.C.D. assembled, finished and annotated sequences; P.C. analysed population genetic and gene expression data; I.J.S. collected the wild sample; I.J.S. and C.J.Y. reared the samples and performed dissections; D.J.R. and A.E.H. built the *cortex* tree; D.J.R. modelled *cortex* structure; M.A.Q. constructed the fosmid library; A.C.D. and N.H. advised on the design of sequencing strategies. The authors declare no competing financial interests.

Corresponding author

I.J.S. (saccheri@liverpool.ac.uk)

Supplementary Information

Supplementary Methods

Supplementary Table 1: Polymorphisms in the *carbonaria* candidate region.

Supplementary Table 2: Polymorphisms in the locus *b-d* region.

Supplementary Table 3: *Carbonaria* morph frequencies in the Manchester area.

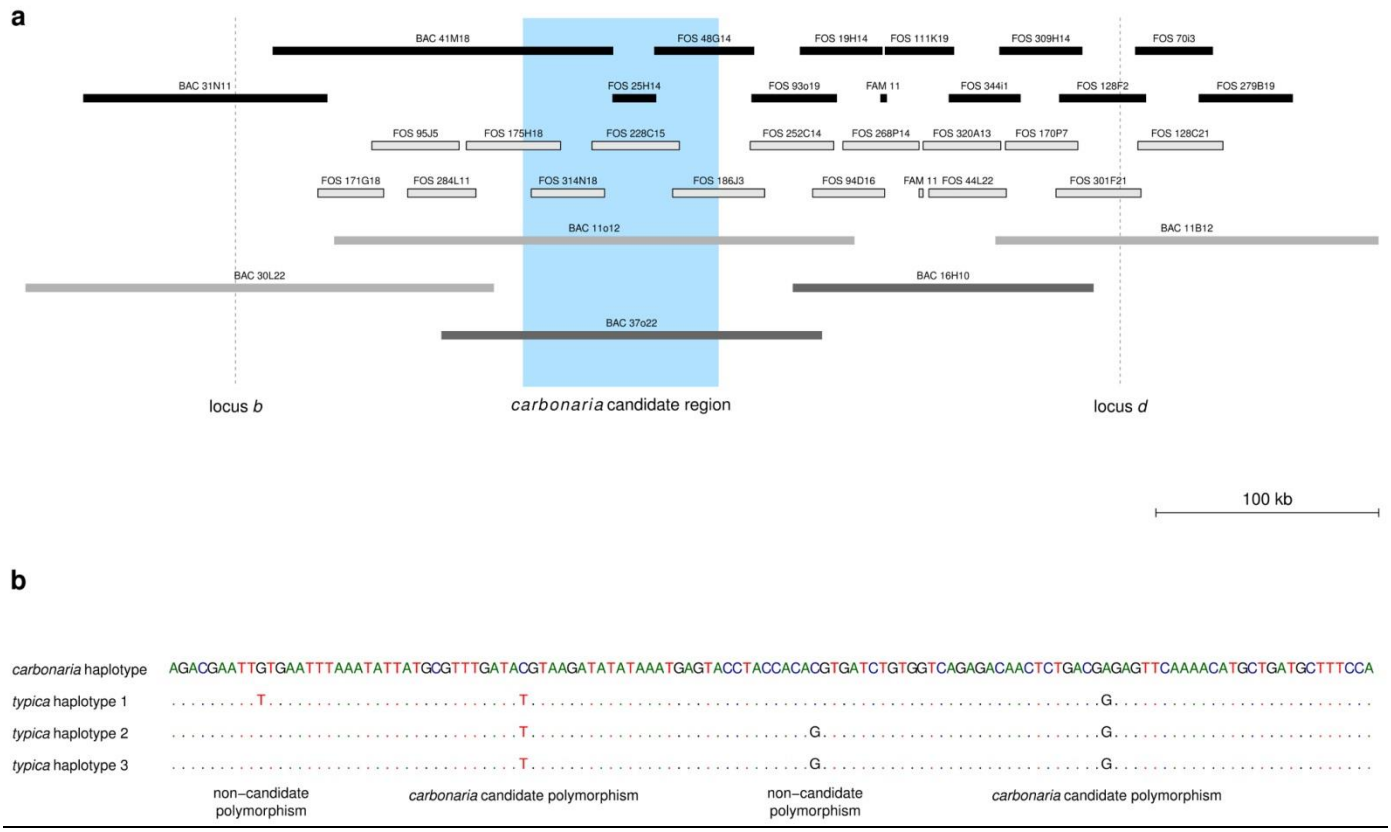
Supplementary Table 4: PCR primers for *cortex*, control genes and candidate genes.

Supplementary Table 5: Sources, including accession numbers, for *cortex* and *Fizzy* family sequences.

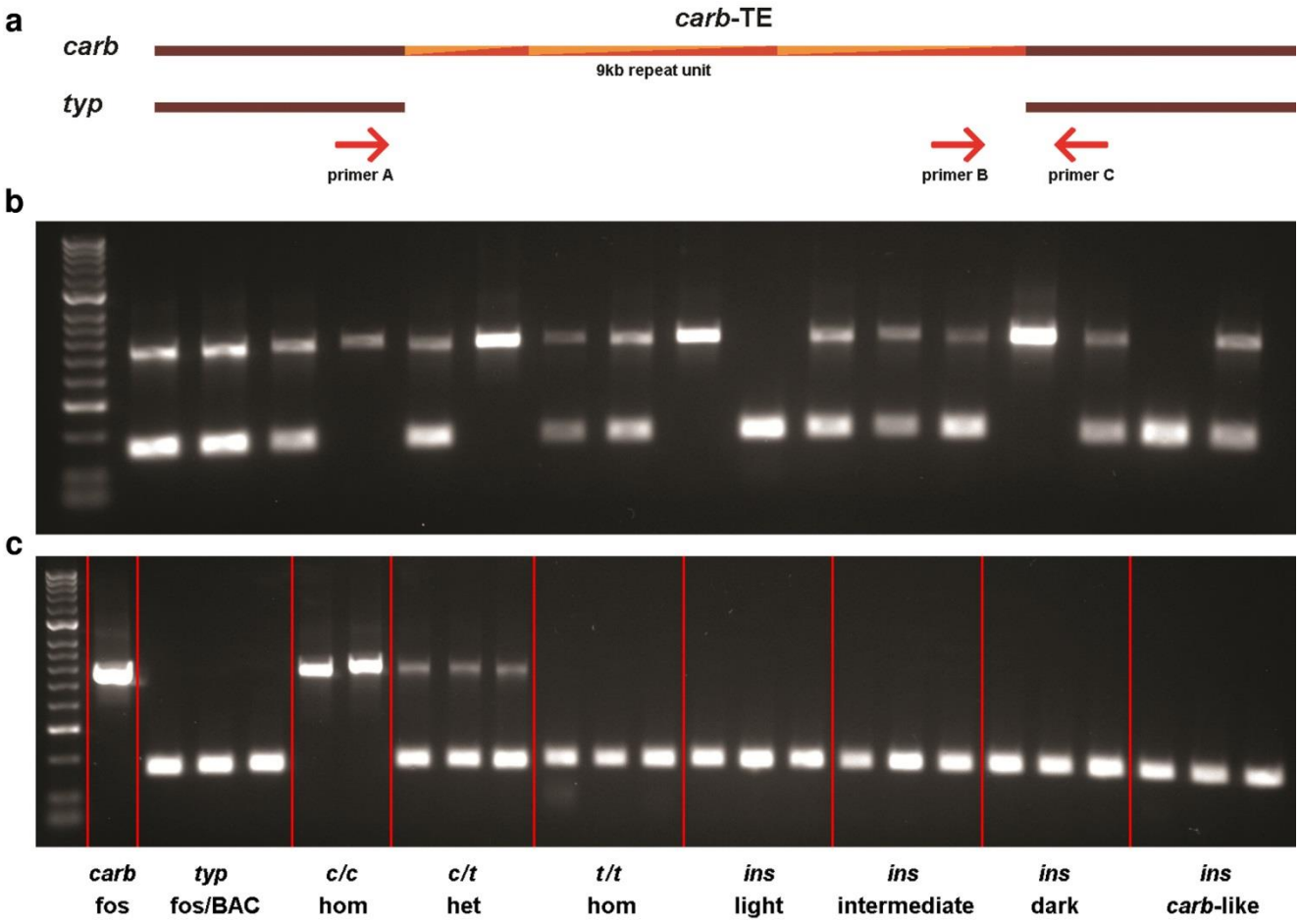
Supplementary Text 1: Sequence alignment of the *carbonaria* and three *typica* haplotypes spanning the '*b-d*' region (illustrated in Extended Data Fig. 1).

Supplementary Text 2: Full-length sequence alignment in aligned FASTA format of cortex proteins and selected homologues. Incompleteness of some sequences at the N-terminus and some uncertainty regarding translation start sites have no impact on the phylogenetic tree presented in Extended Data Fig. 9 since it was calculated using only the propeller domain (Extended Data Fig. 9a).

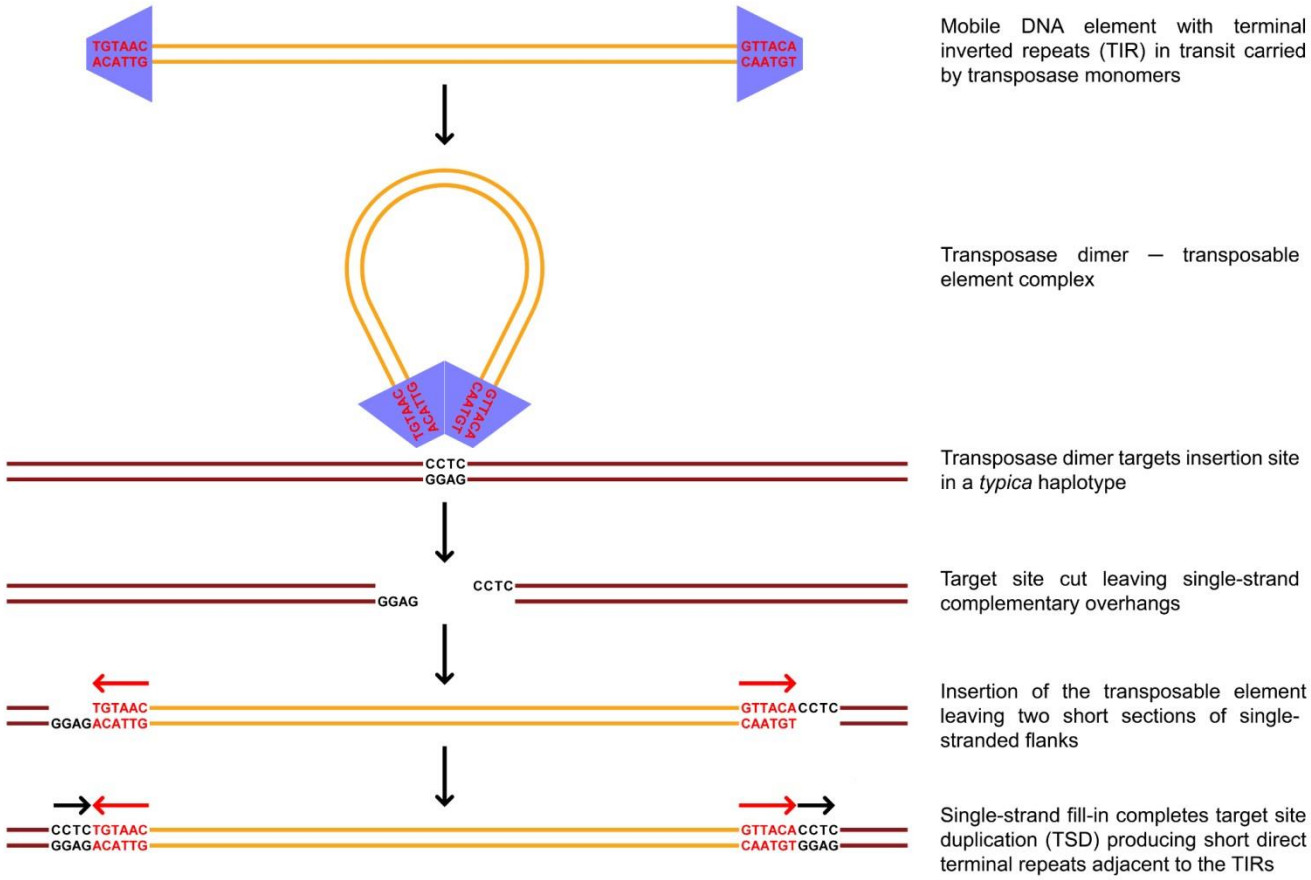
Extended data figures and tables



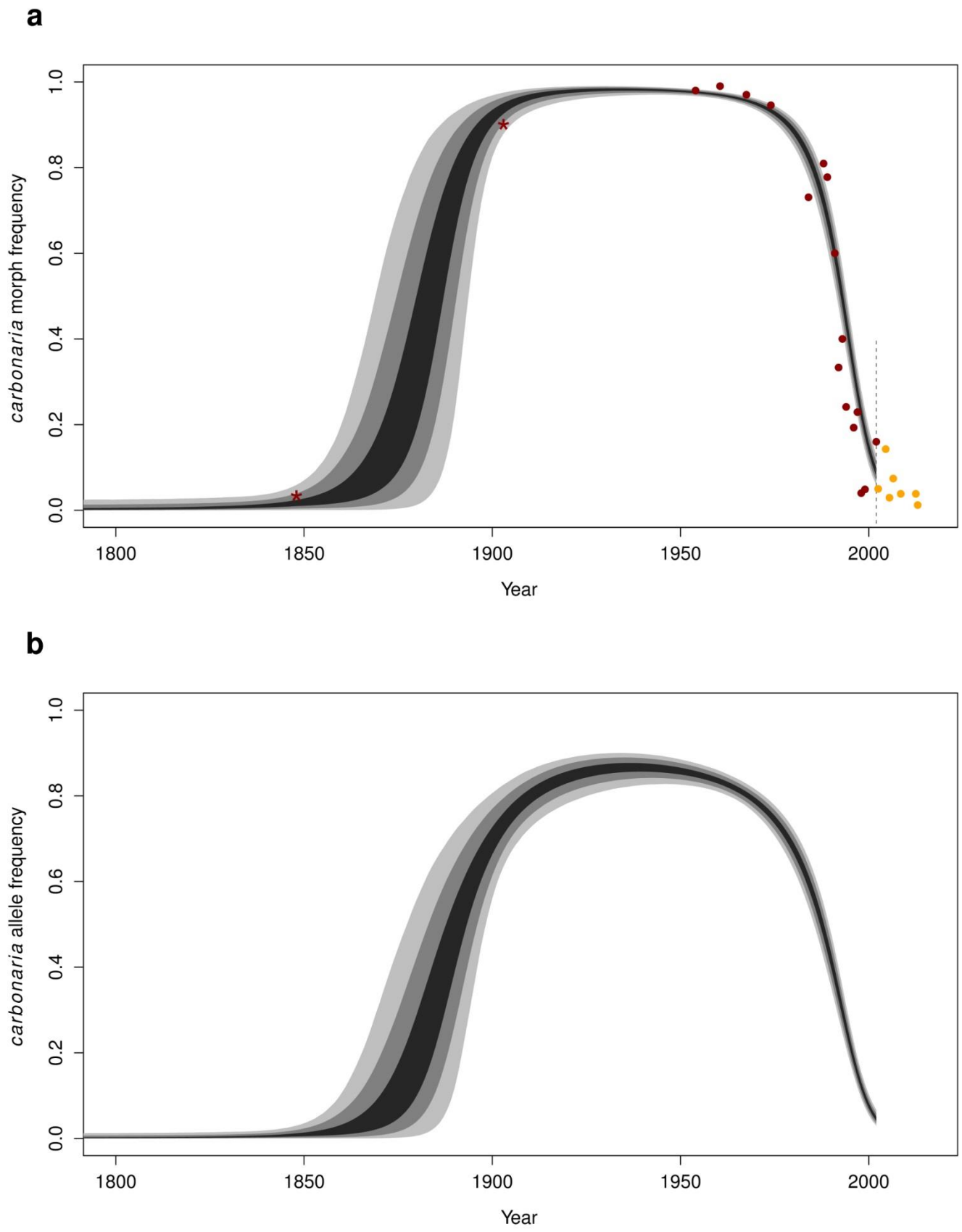
Extended Data Figure 1: BAC and fosmid haplotype tilepaths used to define *carbonaria* candidate polymorphisms. **a**, BAC and fosmid tilepaths of the *carbonaria* haplotype (black bars) and three *typica* haplotypes (different shades of grey). Two small regions not covered by BACs or fosmids were reconstructed using parents and offspring sequences from the same heterozygous family (FAM11). The positions of loci *b* and *d* (cf. Fig. 1) are indicated by the dashed lines, and the *carbonaria* candidate region is highlighted blue. Fosmid 25H14 containing *carb*-TE appears small because it is aligned against *typica* reference sequence which does not include the *carb*-TE. **b**, Alignment of three *typica* haplotypes against the *carbonaria* haplotype for a short section within the *carbonaria* candidate region, showing SNPs (dots are nucleotides identical to the *carbonaria* sequence). Polymorphisms where all three *typica* alleles differ from *carbonaria* were treated as *carbonaria* candidates; polymorphisms where the same allele occurs in *carbonaria* and at least one *typica* were excluded from further consideration.



Extended Data Figure 2: Validation of the 3-primer PCR *carb*-TE genotyping assay in a family and its application in a variety of wild-caught moths. **a**, Schematic alignment of *carbonaria* and *typica* haplotypes showing the position of the three primers (A, B and C, not to scale) used in the same PCR to detect the presence and absence of the 22 kb *carb*-TE. In the presence of *carb*-TE, primers A and C are too far apart to generate a product; the repeat structure of *carb*-TE presents three annealing sites for primer B but only the shortest primer B-C combination is amplified when using 45 s extension (primer sequences are listed in Supplementary Table 1). **b**, *carb*-TE genotypes for father (lane 2), mother (lane 3) and 15 offspring (lanes 4-18); the two brightest bands in the size ladder are 300 bp and 1 kb (lane 1). The parents were full-sibs and known to be heterozygous (*c*/*t*), and therefore expected to generate *c*/*c*, *c*/*t* and *t*/*t* offspring. The larger band (primers B-C) indicates the presence of the *carb*-TE, the smaller band (primers A-C) absence (*typica* allele in this family); heterozygotes have both bands. The individual in lane 15 (135F1-12) is the homozygous male used for wgs. **c**, Presence/absence of *carb*-TE in a *carbonaria* haplotype fosmid clone (lane 2), three different *typica* haplotype clones (lanes 3-5, one fosmid, two BACs), wild *carbonaria* homozygotes (lanes 6-7), wild *carbonaria* heterozygotes (lanes 8-10), *typica* with a flanking haplotype similar to the *carbonaria* haplotype but lacking the *carb*-TE (lanes 11-13), light *insularia* (lanes 14-16), intermediate *insularia* (lanes 17-19), dark *insularia* (lanes 20-22), *carbonaria*-like *insularia* (lanes 23-25).

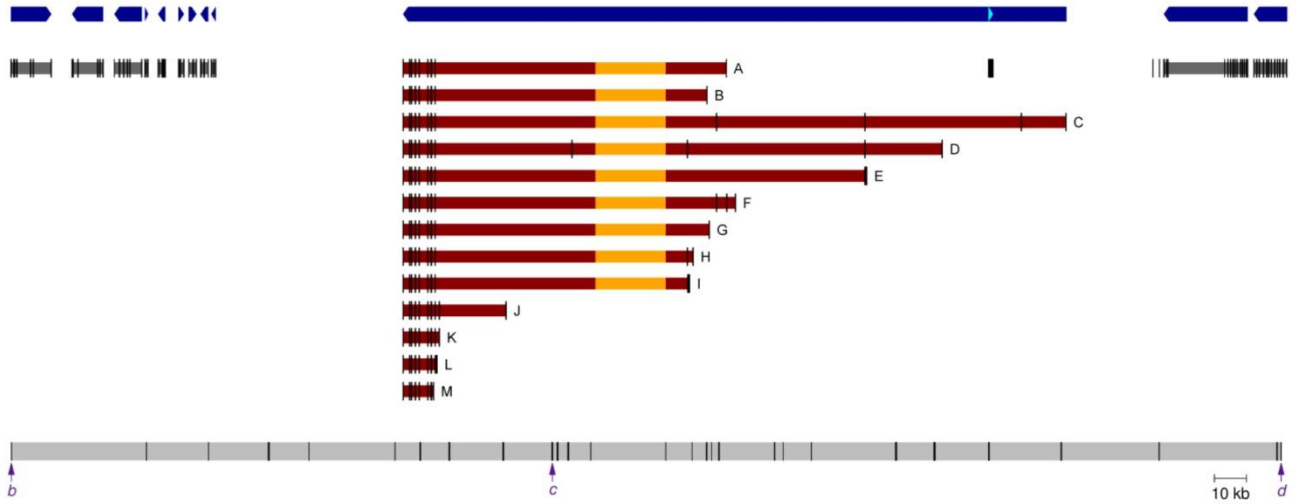


Extended Data Figure 3: Hypothetical reconstruction of the birth of the *carbonaria* allele. Class II non-autonomous DNA transposition is mediated by two transposase monomers linked to terminal inverted repeats (TIR). The monomers form a dimer at the target site which is cleaved to leave short direct repeated overhangs. The transposable element including TIRs is inserted and finally the single-stranded cleaved sites are filled in completing the target site duplication [39]. The unduplicated target site motif (CCTC) is common, possibly ubiquitous, in all non-*carbonaria* (*typica* and *insularia*) haplotypes, but a *typica* ancestor is the more likely given the pattern of haplotype similarities and the presumed prevalence of *typica* haplotypes around 1800.



Extended Data Figure 4: The rise and fall of *carbonaria* in the Manchester area. **a**, Frequency of the *carbonaria* phenotype. **b**, Corresponding frequencies of the *carbonaria* allele. The envelopes show the confidence intervals, 50%, 90%, 99% for the simulated trajectories. Dark-red dots, observations falling within the simulated trajectories; orange dots, additional data collected post-2002 (year during which > 85% of the field sample was collected). Stars indicate likely frequencies where historical data is scarce. Data and sources are listed in Supplementary Table 3.

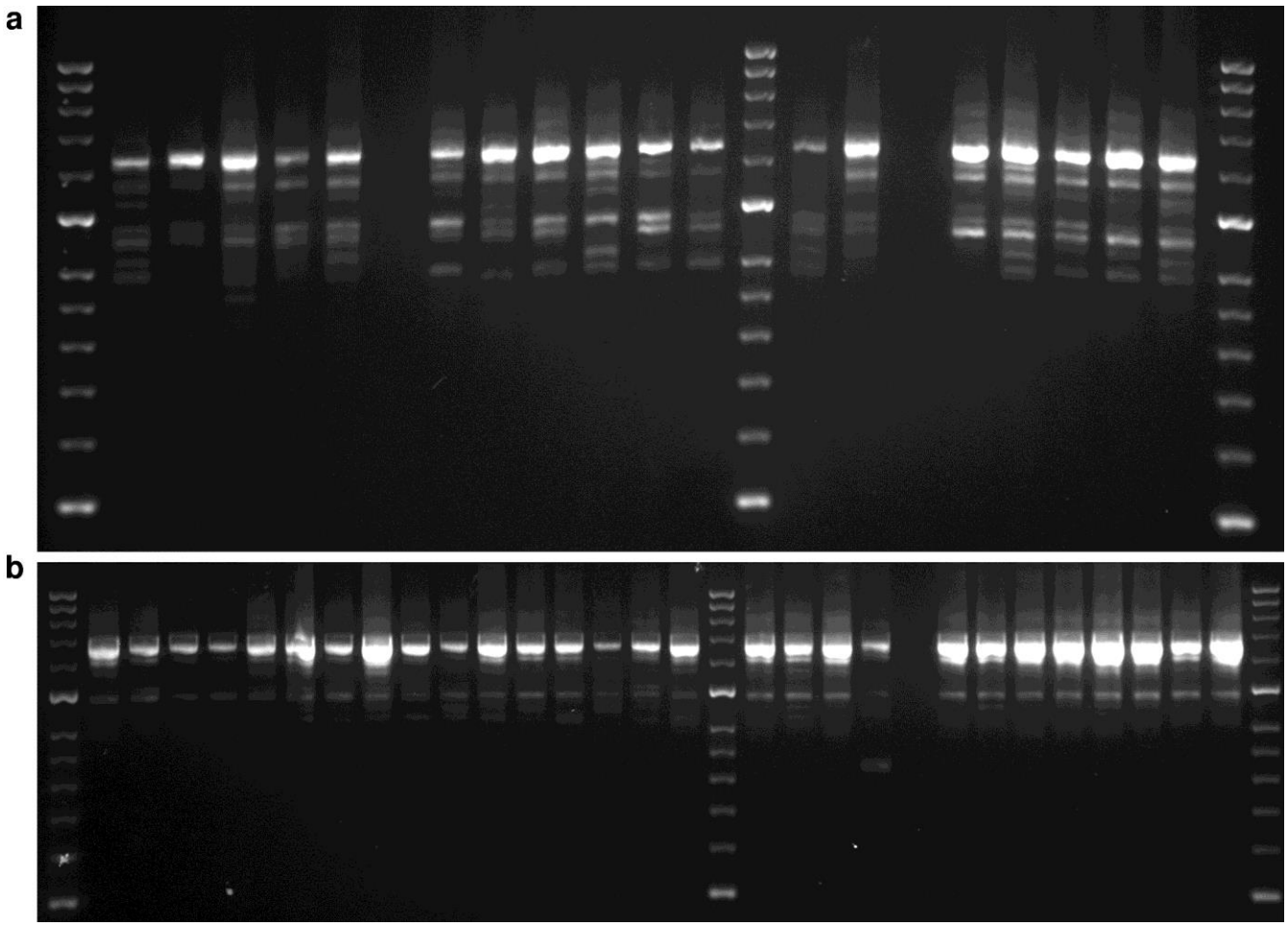
a



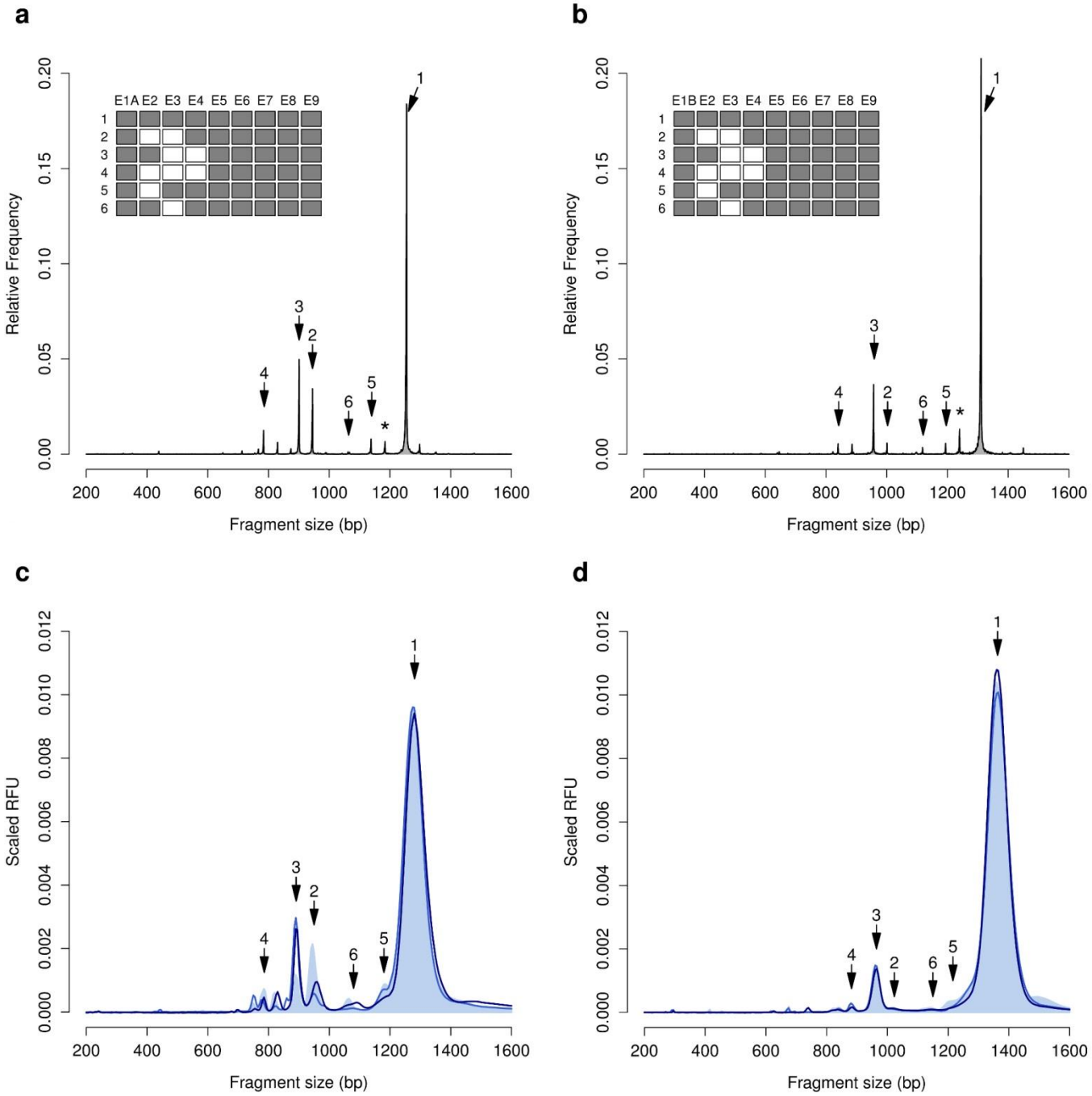
b

Origin	wd/pup	wd	testes	wd	wd	testes	wd	testes	wd	wd	testes	wd	testes
Exon 1	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**
testes	-	++	+	-	+	+	-	-	-	-	++	-	-
ovaries	-	+	-	-	-	-	-	-	-	-	++	+	-
La6_*tt*	+	-	-	-	+	-	-	-	-	-	-	-	-
La6_*cc*	+	-	-	-	+	-	-	-	-	-	-	-	-
Cr2_*tt*	+	++	-	+	-	-	-	-	-	-	-	-	-
Cr2_*ct*	+	++	-	-	-	-	-	-	-	-	+	-	-
Cr4_*tt*	-	+++	-	-	-	-	-	-	-	-	-	+	+
Cr4_*cc*	-	+++	-	-	-	-	-	-	-	-	-	+	-
Cr6_*tt*	-	+	+	-	-	-	-	-	-	-	+	-	-
Cr6_*cc*	-	+	+	-	-	-	-	-	-	-	+	-	-
Pu2_*tt*	+	+	-	-	+	-	-	-	-	-	-	+	-
Pu2_*ct*	+	+	-	-	-	-	-	-	-	-	-	+	-
PDP_*tt*	+	++	-	-	-	-	-	-	-	-	-	-	-
PDP_*ct*	+	++	-	-	-	-	-	-	-	-	-	-	-

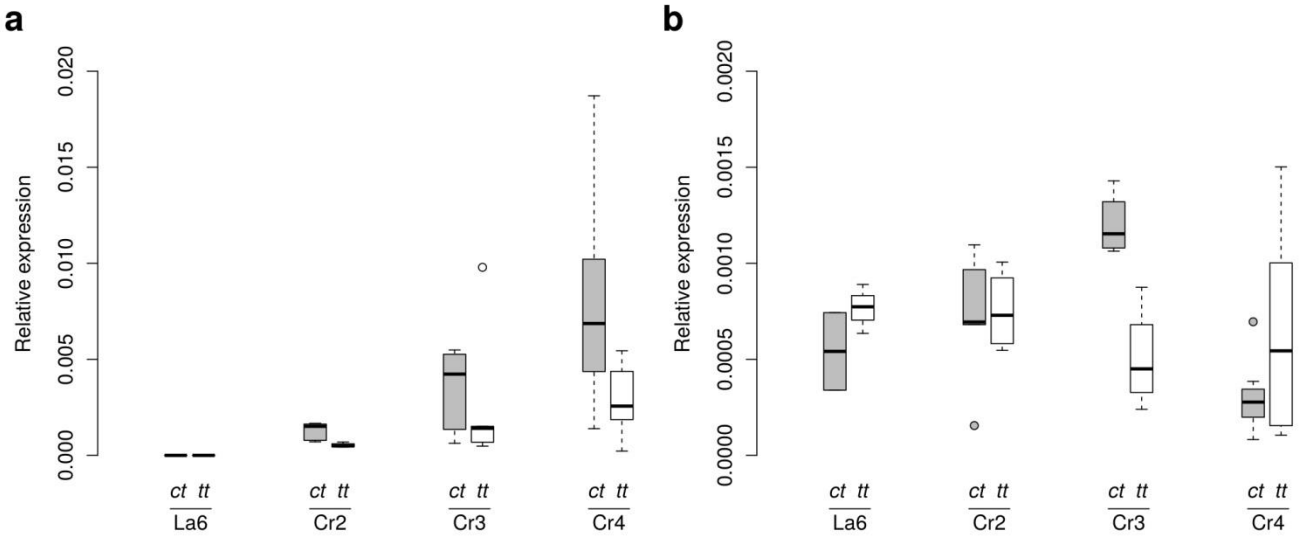
Extended Data Figure 5: a, Illustration of *cortex* exon structure indicating the positions of thirteen alternative transcription starts and subsequent exons relative to the flanking genes in the *b-d* region (position of *carb*-TE indicated by orange bar). **b**, Expression of different starting position *cortex* transcripts. End-point reverse transcription PCR with reduced cycles (35) was used to exclude transcripts with negligible dosage. Amplicon intensities are scaled between + (faint but visible) and +++ (strong PCR product). Negative PCRs represent expression below the detection threshold; this may even occur in 'origin' tissue types (wing disc/pupa/testes) in which the alternative starts were discovered owing to the fact that 5' RACE used ~20 times the amount of RNA template relative to the standard cDNA synthesis for the 35 cycle end-point PCRs. Ovaries were not used for 5' RACE which may have caused gonad expression bias towards testes. Test tissues are 6th instar larvae gonads and wing discs at different developmental stages (abbreviations as in Fig. 3).



Extended Data Figure 6: Examples of *cortex* splice variation pattern in *typica* and *carbonaria* developing wing discs. End-point PCR on wing disc cDNA amplified with primers in the first and last exons (E1-E9), with *typica* individuals to the left of the central ladder (the two brightest bands in the size ladder are 300 bp and 1 kb) and *carbonaria* individuals (all *carbonaria-typica* heterozygotes) to the right of the central ladder. **a**, Exon 1A variants in Cr2 stage. **b**, Exon 1B variants in Cr4 stage. (See Fig. 3 for stage abbreviations.)



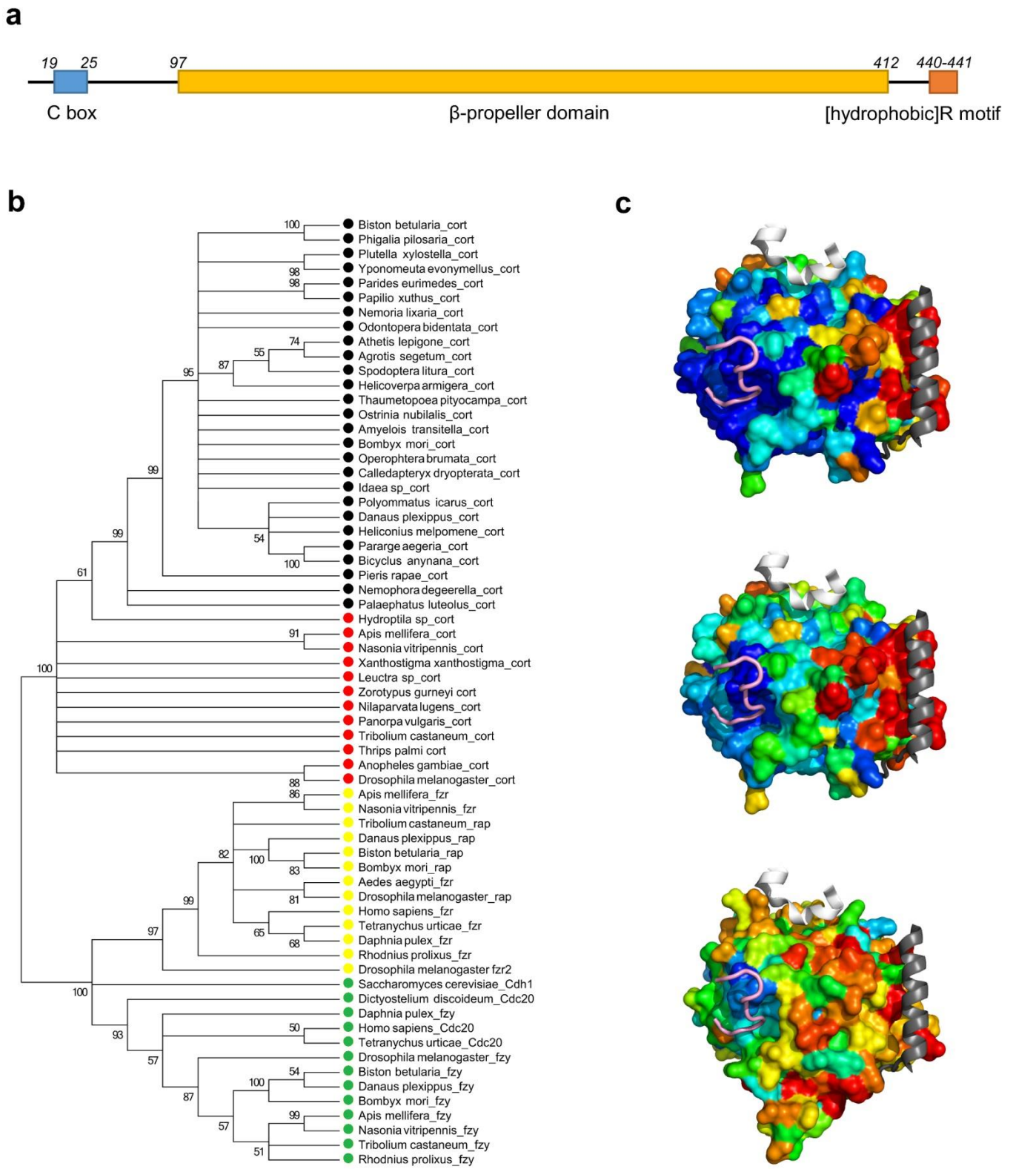
Extended Data Figure 7: Exonic structure and size distributions of *cortex* splice variants amplified by end-point RT-PCR with primers in exon 1A or 1B and exon 9. Size distributions of the PacBio reads are displayed for the two alternative first exons 1A (**a**) and 1B (**b**) of *cortex*. **c** and **d**, Comparison of *carbonaria*-locus genotypes (*t/t* pale blue fill, *c/t* light blue line, *c/c* dark blue line) measured with Fragment Analyzer. Relative Fluorescence Units (RFU) were averaged across individuals for fragments amplified with E1A-E9 (**c**) or E1B-E9 (**d**) primers. Prior to averaging RFUs were standardized so that the total fluorescence (area under the curve) per individual scales to 1. Arrows with the same number denote either similar exonic structure (E1A vs E1B variants) or fragment identity between the two sources of data (PacBio reads and Fragment Analyzer). Exonic structure of the six main splice variants is represented in matrices (**a**, **b**), in which white cells represent skipped exons in a splice variant (* indicates full transcript in which the first 71bp of exon 6 are missing). [Apparent differences among melanic and non-melanic for 1A #2 and #3 splice variants were not consistent among families].



c

	E7-E9 Fig. 3a			E1B-E3 Fig. 3c			E1A-E3 Fig. 3d			E1B-E3 ED Fig. 8a		E1A-E3 ED Fig. 8b	
Stage	*cc*	*ct*	*tt*	*cc*	*ct*	*tt*	*cc*	*ct*	*tt*	*ct*	*tt*	*ct*	*tt*
La6	6	12	20	6	3	7	6	5	10	2	3	2	3
Cr2	0	8	11	0	12	8	0	12	8	5	4	5	4
Cr3	0	6	8	10	9	1	9	9	4	5	6	5	6
Cr4	5	22	22	4	7	5	5	7	5	12	13	12	13
Cr5	0	4	4	5	6	4	5	6	4	0	0	0	0
Cr6	0	5	5	0	0	0	0	0	0	0	0	0	0
Pu2	0	4	5	0	0	0	0	0	0	0	0	0	0
Pu4	0	3	7	0	0	0	0	0	0	0	0	0	0
Pu6	0	4	4	0	0	0	0	0	0	0	0	0	0
Pu8	0	5	3	0	0	0	0	0	0	0	0	0	0
PDP	0	3	17	0	0	0	0	0	0	0	0	0	0

Extended Data Figure 8: Relative expression of *cortex* full transcript in developing wing discs, comparing *ct* heterozygotes with *tt* homozygotes produced from *ct* x *tt* crosses (starting with exon 1B [**a**] or exon 1A [**b**]). Genotypes differ significantly for 1B full transcript (P = 0.001), whereas genotypes do not differ for 1A full transcript (P > 0.5). Note the differing y-axis scales. **c**, Sample sizes for *cortex* RT-qPCR experiments by wing disc developmental stage and *carbonaria*-locus genotypes. (La6: 6th instar larvae, Cr2: day 2 crawler, Pu2: day 2 pupae, PDP: post diapause pupae.)



Extended Data Figure 9: a, Schematic illustration, not to scale, of molecular features of *B. betularia* cortex protein sequence. **b**, Bootstrapped Maximum Likelihood consensus tree calculated with MEGA 6 of Fizzy/cortex derived from the propeller domain of the alignment in Supplementary Text 2. Branches are collapsed where partitions were reproduced in less than half of bootstrap replicates. Major groups containing lepidopteran cortex (black circles), non-lepidopteran cortex (red circles), Fizzy-related/rap (yellow circles) or Fizzy/Cdc20/Cdh1 proteins (green circles), are similarly unequivocally defined in trees obtained by Neighbour Joining or Maximum Parsimony methods (not shown). **c**, 3D protein sequence conservation mapping of: upper panel, lepidopteran cortex sequences onto a homology model of *B. betularia* cortex; middle panel, all cortex sequences onto the same *B. betularia* model; lower panel, non-lepidopteran cortex sequences onto a model of *D. melanogaster* cortex. Molecular surfaces are shown in PyMOL using a spectrum from high conservation (blue) to low (red). The mapping reveals the shared presence of a presumed inter-blade D box-like degron-binding site (pink segment is superimposed D box-mimicking sequence from the structures of human APC/C [PDB 4ui9] [40]). In contrast, there is much weaker conservation of surface regions corresponding to facial KEN box or helical specificity-determinant sites (white and grey regions, respectively, from the same structure), suggesting that cortex proteins lack these functionalities. Note that the greater sequence variability in the non-lepidopteran set leads to lower overall sequence conservation (lower panel) but that overall patterns in all panels are similar.

Extended Data Table 1: Predicted functionality of *B. betularia* cortex isoforms (starting with exon 1A or 1B).

		Feature known in Cdh1/Cdc20 (function) and its potential conservation					
		Binding to APC/C			Binding to degrons (see Ext. Dat. Fig. 9)		
Isoform[1]	Length (residues)	C-box: DRFVVPR (binds Apc8 subunit of APC/C)	Segments 2 & 4 (bind Apc1)	[hydrophobic]R C-terminus (binds Apc3)	Inter-blade recognition site for LxExxxN degron	Facial recognition of KEN-box degron	Recognition of helical specificity determinant
1A	441	✔	✖	✔	✔	✖	✖
1B	407	✖	✖	✔	✔	✖	✖
2A	291	✖	✖	✔	✖[2]	✖	✖
2B	291	✖	✖	✔	✖[2]	✖	✖
3A	323	✔	✖	✔	✖	✖	✖
3B	289	✖	✖	✔	✖	✖	✖
4A	284	✖	✖	✔	✖	✖	✖
4B	270	✖	✖	✔	✖	✖	✖
5A	402	✖	✖	✔	✔	✖	✖
5B	270	✖	✖	✔	✖	✖	✖
6A	291	✖	✖	✔	✖[2]	✖	✖
6B	291	✖	✖	✔	✖[2]	✖	✖

[1] Isoforms as defined in Extended Data Fig. 7.

[2] Since the region lost from the propeller fold constitutes approximately a single blade, it is possible that these, and only these, truncated-propeller forms may still fold stably.

References

1 Cook, L. M. The rise and fall of the *carbonaria* form of the peppered moth. *Q. Rev. Biol.* **78**, 399-417 (2003).
2 Van't Hof, A. E., Edmonds, N., Dalíková, M., Marec, F. & Saccheri, I. J. Industrial melanism in British peppered moths has a singular and recent mutational origin. *Science* **332**, 958-960 (2011).
3 Brookfield, J. F. Y. Evolutionary genetics: mobile DNAs as sources of adaptive change? *Curr. Biol.* **14**, R344-R345 (2004).
4 Barrett, R. D. H. & Hoekstra, H. E. Molecular spandrels: tests of adaptation at the genetic level. *Nat Rev Genet* **12**, 767-780 (2011).
5 Nadeau, N. J. & Jiggins, C. D. A golden age for evolutionary genetics? Genomic studies of adaptation in natural populations. *Trends Genet.* **26**, 484-492 (2010).
6 Martin, A. & Orgogozo, V. The loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. *Evolution* **67**, 1235-1250 (2013).
7 Stern, D. L. The genetic causes of convergent evolution. *Nat Rev Genet* **14**, 751-764 (2013).
8 Savolainen, O., Lascoux, M. & Merila, J. Ecological genomics of local adaptation. *Nat Rev Genet* **14**, 807-820 (2013).
9 Hoekstra, H. E. & Coyne, J. A. The locus of evolution: evo devo and the genetics of adaptation. *Evolution* **61**, 995-1016 (2007).
10 Cook, L. M. & Saccheri, I. J. The peppered moth and industrial melanism: evolution of a natural selection case study. *Heredity* **110**, 207-212 (2013).
11 Chu, T., Henrion, G., Haegeli, V. & Strickland, S. *Cortex*, a Drosophila gene required to complete oocyte meiosis, is a member of the Cdc20/fizzy protein family. *Genesis* **29**, 141-152 (2001).
12 Saccheri, I. J., Rousset, F., Watts, P. C., Brakefield, P. M. & Cook, L. M. Selection and gene flow along a diminishing cline of melanic peppered moths. *Proc. Natl. Acad. Sci. USA* **105**, 16212-16217 (2008).
13 Clarke, C. A. *Biston betularia*, obligate f. *insularia* indistinguishable from f. *carbonaria* (Geometridae). *Journal of the Lepidopterists' Society* **33**, 60-64 (1979).
14 Lees, D. R. & Creed, E. R. Genetics of *insularia* forms of peppered moth, *Biston betularia*. *Heredity* **39**, 67-73 (1977).
15 Kim, Y. & Nielsen, R. Linkage disequilibrium as a signature of selective sweeps. *Genetics* **167**, 1513-1524 (2004).
16 Cook, L. M., Sutton, S. L. & Crawford, T. J. Melanic moth frequencies in Yorkshire, an old English industrial hot spot. *J. Hered.* **96**, 522-528 (2005).
17 Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9**, 397-405 (2008).
18 He, J. *et al.* Insights into degron recognition by APC/C coactivators from the structure of an Acm1-Cdh1 complex. *Mol. Cell* **50**, 649-660 (2013).
19 Whitfield, Z. J., Chisholm, J., Hawley, R. S. & Orr-Weaver, T. L. A meiosis-specific form of the APC/C promotes the oocyte-to-embryo transition by decreasing levels of the polo kinase inhibitor matrimony. *PLoS Biol* **11**, e1001648 (2013).
20 Nadeau, N. J. *et al.* A major gene controls mimicry and crypsis in butterflies and moths. *Nature* **in press** (2015).
21 Ito, K. *et al.* Mapping and recombination analysis of two moth colour mutations, *Black moth* and *Wild wing spot*, in the silkworm *Bombyx mori*. *Heredity* **116**, 52-59 (2016).
22 González, J., Karasov, T. L., Messer, P. W. & Petrov, D. A. Genome-wide patterns of adaptation to temperate environments associated with transposable elements in *Drosophila*. *PLoS Genet.* **6**, e1000905 (2010).
23 Schlenke, T. A. & Begun, D. J. Strong selective sweep associated with a transposon insertion in *Drosophila simulans*. *Proc. Natl. Acad. Sci. USA* **101**, 1626-1631 (2004).
24 Schrader, L. *et al.* Transposable element islands facilitate adaptation to novel environments in an invasive species. *Nat Commun* **5** (2014).
25 Casacuberta, E. & González, J. The impact of transposable elements in environmental adaptation. *Mol. Ecol.* **22**, 1503-1517 (2013).
26 Koga, A., Iida, A., Hori, H., Shimada, A. & Shima, A. Vertebrate DNA transposon as a natural mutator: the medaka fish *Tol2* element contributes to genetic variation without recognizable traces. *Mol. Biol. Evol.* **23**, 1414-1419 (2006).
27 Van't Hof, A. E. *et al.* Linkage map of the peppered moth, *Biston betularia* (Lepidoptera, Geometridae): a model of industrial melanism. *Heredity* **110**, 283-293 (2013).
28 Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465-W467 (2005).
29 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotech* **29**, 644-652 (2011).
30 Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179-181 (2012).
31 Agapow, P.-M. & Burt, A. Indices of multilocus linkage disequilibrium. *Mol. Ecol. Notes* **1**, 101-102 (2001).

32 Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet* **8**, e1002453 (2012).
33 Baxter, S. W. *et al.* Genomic hotspots for adaptation: the population genetics of Müllerian mimicry in the *Heliconius melpomene* clade. *PLoS Genet.* **6**, e1000794 (2010).
34 Reed, R. D., McMillan, W. O. & Nagy, L. M. Gene expression underlying adaptive variation in *Heliconius* wing patterns: non-modular regulation of overlapping cinnabar and vermilion prepatterns. *Proc. R. Soc. B* **275**, 37-46 (2008).
35 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780 (2013).
36 Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725-2729 (2013).
37 Šali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779-815 (1993).
38 Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* **38**, W529-W533 (2010).
39 Muñoz-López, M. & García-Pérez, J. L. DNA transposons: nature and applications in genomics. *Curr. Genomics* **11**, 115-128 (2010).
40 Chang, L., Zhang, Z., Yang, J., McLaughlin, S. H. & Barford, D. Atomic structure of the APC/C and its mechanism of protein ubiquitination. *Nature* **522**, 450-454 (2015).